Published in: Konopka, Marek/Kubczak, Jacqueline/Mair, Christian/Šticha, František/
Waßner, Ulrich Hermann (eds.): Grammatik und Korpora 2009. Dritte Internationale
Konferenz. Mannheim, 22.-24.9.2009. - Tübingen: Narr, 2011. pp. 451-469.
(Korpuslinguistik und interdisziplinäre Perspektiven auf Sprache 1)

CYRIL BELICA / MARC KUPIETZ / ANDREAS WITT / HARALD LÜNGEN

The Morphosyntactic Annotation of DeReKo: Interpretation, Opportunities, and Pitfalls

Abstract

The paper discusses from various angles the morphosyntactic annotation of DeReKo, the Archive of General Reference Corpora of Contemporary Written German at the Institut für Deutsche Sprache (IDS), Mannheim. The paper is divided into two parts. The first part covers the practical and technical aspects of this endeavor. We present results from a recent evaluation of tools for the annotation of German text resources that have been applied to DeReKo. These tools include commercial products, especially Xerox' Finite State Tools and the Machinese products developed by the Finnish company Connexor Oy, as well as software for which academic licenses are available free of charge for academic institutions, e.g. Helmut Schmid's Tree Tagger. The second part focuses on the linguistic interpretability of the corpus annotations and more general methodological considerations concerning scientifically sound empirical linguistic research. The main challenge here is that unlike the texts themselves, the morphosyntactic annotations of DeReKo do not have the status of observed data; instead they constitute a theory and implementation-dependent interpretation. In addition, because of the enormous size of DeReKo, a systematic manual verification of the automatic annotations is not feasible. In consequence, the expected degree of inaccuracy is very high, particularly wherever linguistically challenging phenomena, such as lexical or grammatical variation, are concerned. Given these facts, a researcher using the annotations blindly will run the risk of not actually studying the language but rather the annotation tool or the theory behind it. The paper gives an overview of possible pitfalls and ways to circumvent them and discusses the opportunities offered by using annotations in corpus-based and corpus-driven grammatical research against the background of a scientifically sound methodology.

1. Introduction

This paper is inspired by a recent corpus annotation venture at the Institut für Deutsche Sprache (IDS) in Mannheim, Germany. The focus of the paper is on two separate yet related topics. On the one hand, a brief chronological overview of the linguistic annotations of DeReKo,[1] the Archive of General Refer-

[1] The name was inspired by a research project running from 1999 to 2001 in co-operation with the universities of Stuttgart and Tübingen.

ence Corpora of Contemporary Written German at the IDS, is presented, followed by a more detailed and more technical account of the planning, decision-making, deployment, and evaluation processes involved in the current annotation phase. Among other perspectives, a separate short discussion of the inter-tagger agreement between (1) Xerox' Finite State Tools, (2) the Connexor Oy's Machinese products, and (3) Helmut Schmid's Tree Tagger – as observed in the 3.75 billion sized DeReKo – is included.

On the other hand, the paper examines the potential methodological difficulties encountered when linguists include the annotated DeReKo (or any other very large annotated corpus) in their scientific research context. We conjecture that whenever linguistically challenging phenomena such as, in particular, language variation are studied, the observed annotation inaccuracy might prove to be worrisomely high, and, moreover, biased in a systematic way at that. We argue that a linguist trusting the annotations blindly would run the risk of not actually exploring the language captured in the corpus but that he or she would rather be detailing the annotation tool or the linguistic theory behind it instead. Finally, we discuss how linguists might want to circumvent such pitfall traps in order to approach very large annotated corpora in a scientifically sound way.

2. Tagging the IDS-corpora

The corpora of contemporary written German at the IDS, since 2004 called *Deutsches Referenzkorpus* – DeReKo, are one of the major resources worldwide for the study of the German language (Kupietz / Keibel 2009). The first step towards providing these corpora with access to linguistic annotation was already made in 1993, when a new version of the Corpus Search, Management and Analysis System COSMAS (IDS 1991-2009) was planned specifically in order to be capable of handling multi-layer annotations. Subsequently, the corpora were tagged several times, most notably in 1995 with the Logos Tagger and in 1999 with the Gertwol Tagger (Koskenniemi / Haapalainen 1996).

As shown in Table 1, the current annotation initiative of DeReKo also reaches back to 2002 when its start and co-ordination with the COSMAS project was incorporated into the IDS research plan 2003-2008.

2002-10	new annotation initiative incorporated in the IDS Research Plan 2003-2008
2007-01	COSMAS II confirms to support multiple stand-off annotation search by 2008
2007-10	process model for the annotation of DeReKo
2007-12	catalog of criteria for tagging tools
2008-04	start of market analysis
2008-07	25 tools → shortlist of 9
2008-08	request for evaluation versions
2008-09	start of in-depth expert study
2008-12	→ 3 tools recommended
2008-12	first annotation attempts
2009-02	filter and collation scripts developed to support XML stand-off annotations
2009-07	after 6 cpu years: 3.5 TB of annotation data produced
2009-08	DeReKo-2009-II released with full TreeTagger and Connexor annotations and partial Xerox annotations

Table 1: History of the current annotation initiative for DeReKo

2.1 Selection of tools

The deployment of the current annotation initiative was launched in late 2007, when a first coarse process model for the annotation of DeReKo with the following cornerstones was developed:

1) do not rely on judgements of a single tagger, i.e., provide multiple concurring annotations that result from different tools;
2) for every tagger include as many concurrent interpretations of each linguistic phenomenon as possible;
3) use different types of taggers to avoid systematic biases;
4) consider annotating multiple linguistic levels if appropriate tools are available;
5) invite an external expert panel to (1) carry out a market analysis in order to put up a shortlist of suitable taggers and to (2) conduct an in-depth study of the short-listed taggers in order to arrive at a final recommendation;

6) after completing the annotation phase, evaluate each annotation layer with respect to fitness for their particular intended use in linguistic research.

The next step was then to find and engage external experts in the field of computational linguistics, part-of-speech tagging, and other levels of automatic linguistic annotation to carry out an independent study, resulting in a shortlist (phase 1) and, finally, in a list of recommended tools (phase 2). Together with these experts, at first, a catalogue of linguistic, organizational, technical, and economic selection criteria was developed that can be summarized as follows:

- *linguistic:* reliability, precision, recall, disambiguation, self-assessment, tagset compatibility, types of analysis (POS, dependencies, ...), extensibility;
- *organizational:* long term perspectives, sustainability, "also applied by";
- *technical:* supported platforms, adaptability, maintainability, robustness, i/o-formats, resource requirements;
- *economic:* licensing options, license restrictions, pricing.

The subsequent market analysis was started in April 2008. Its result was a brief evaluation of about 25 tools according to the selection criteria and a shortlist of nine tools, recommended for closer investigation in the second phase of the expert study:

- GERTWOL (Lingsoft Oy)
- Machinese Tools (Connexor Oy)
- SMOR (Stuttgart University)
- TAGH (Thomas Hanneforth)
- TNT (Thorsten Brants)
- TreeTagger (Stuttgart University)
- Unsupos (Leipzig University)
- WMTrans (Canoo AG)
- Xerox FST Linguistic Suite (Xerox Company)

After an internal review of the results of the study, we decided to proceed as recommended by the expert panel. Evaluation versions of the nine tools were requested and – if granted – assessed more thoroughly in the in-depth study carried out by the experts between September and December 2008. This phase resulted in a final recommendation of three taggers:

- Machinese Tools from Connexor Oy,
- TreeTagger from Stuttgart University, and
- Xerox FST Linguistic Suite.

They are described in more detail below. Again, following the expert recommendations, we eventually decided to acquire the necessary licenses for these tools.

For our purposes, the TreeTagger was found to be the best tagger available free of charge. It employs a statistical tagging method in which transition probabilities are estimated by decision trees, hence its name (Schmid 1994). It provides disambiguated morphological and POS information in the form of STTS tags (Schiller et al. 1999). Its parameter files can be updated, i.e., the TreeTagger can be re-trained with one's own correctly tagged material. Moreover, it is continuously and actively being developed by its author Helmut Schmid at Stuttgart University.

The commercial Xerox FST Linguistic Suite from Xerox Inc., USA / France, provides very accurate tagging with rule-based POS disambiguation but no disambiguation of morphological tags. The tag set used is similar but not identical to STTS. The acquired license does not allow for a publication of tagged corpora (presumably to avoid the danger of reverse engineering).

The third recommendation consisted of the commercial products Machinese Phrase Tagger and Machinese Syntax from Connexor Oy, Helsinki. Machinese Phrase Tagger has the same functionality as Xerox FST, and, as far as we could infer from the description of these commercial products, both tools use similar techniques for this task. Machinese Syntax was the only tool tested that provides actual syntactic structures. Its analysis is based on the Functional Dependency Grammar (Tapanainen / Järvinen 1997), including, amongst other things, disambiguated POS and morphological tags. The acquired license does not allow for a publication of tagged corpora, i.e., its restrictions are comparable to those of newspaper text in DeReKo, from which only short passages may be quoted.

In the annotation cycle outlined in the following section, we only used the morphological and part-of-speech analysis components of these three tools.

2.2 The tagging process

To be able to apply the tools to the XCES-encoded DeReKo data, filters had to be developed first to mask out text that should not be annotated (e.g. metadata). In addition, for the output of the tools, postprocessors had to be developed to produce a uniform stand-off XML format. The latter was not trivial because only the Connexor tool was able to give information about character offsets of the analyzed surface forms in the original text.

Once the pre-processing and post-processing scripts worked sufficiently well on a sample of DeReKo, the process of annotating the whole DeReKo started in March 2009. As there were no versions of the tools for our default platform Solaris x86 and our provisional tests had shown that part-of-speech tagging with disambiguation and morphological tags was quite time-consuming – especially in the case of the Xerox tool –, a new (and thus untested) Linux machine with 32 cores (AMD Opteron 8356 processors at 2.3 GHz) and 256 GB RAM was borrowed from another IDS project.

The processing of DeReKo's approx. 350 XCES-files with up to 2 GB each had to be interrupted and restarted several times because new offset-linking problems with implications for the pre-processor were detected or because of hardware problems with the untested machine. In July 2009, after about 6 CPU-years (taking account of all restarts), the annotations with the TreeTagger and the Connexor tool were finished and the Xerox annotation was suspended at about 60 % to have time for a first evaluation before the DeReKo-2009-II release scheduled for August.

By then, the size of stand-off XML annotation data containing all part-of-speech and morphological analyses provided by the tools totaled 3.5 TB and, in first iteration, cornerstones 1-5 (see Section 2.1, Selection of tools) were put into action.

3. Analysis of tagging results

3.1 Methodology

In order to obtain a first impression of the reliability and usability of the tools for linguistic tasks without getting lost in 3.5 TB of annotation data, we decided to start our analysis with a superficial comparison of the outputs of the three taggers deliberately ignoring everything but POS information and everything that was not easily comparable.

3.1.1 Tag sets

To be able to perform the comparison, we first had to define a basic tag set and mappings from the original tag sets. The result was the tag set B_9 shown in Table 2 with nine part-of-speech categories. B_9 is a true subset of the Connexor base tag set, leaving out its categories for *interjection* (INTERJ) and *subordinating conjunction* (CS), which had no consistent correspondents in neither the TreeTagger nor the Xerox tag set. From the Xerox and TreeTagger tag sets the tags DATE, FM, KOKOM, KOUI, KOUS, PTKVZ, TRUNC, and XY (see Schiller et al. 1999) were not consistently mappable on common tags. According to our conservative approach, whenever a non-mappable tag was encountered in a comparison of decisions, the comparison was ignored and not counted.

Mainly in order to obtain a rough idea of how big the influence of the tag set granularity on the comparison of results was, we also defined a tentative more fine-grained base tag set B_{26} with 26 different categories derived from the STTS including some disjunctions like PRELS / PRELAT and PDS / PDAT. As the mapping was not straightforward and is in need of further inspection, results based on it will be reported in parentheses.

3.1.2 Corpus

The sample corpus we conducted the comparison on was the DeReKo-based virtual corpus (cf. Kupietz / Keibel 2009) POScomp09a with 370 million words in 1.7 million texts from mainly German newspapers from 1997 to 2009.[2]

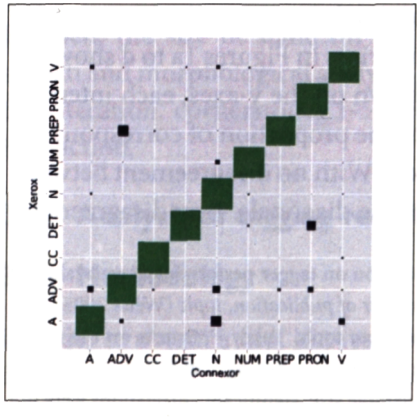
Based on the coarse base tag set B_9, we examined the average POS tag correspondence of pairs of tools. The confusion matrices in Figures 1a to d show the results. Based on the decision by the tool shown on the y-axes, each intersection point has a square with a size relative to the proportion of corresponding classifications by the tool shown on the x-axes. With no disagreement between the tools, the squares would only appear on the diagonals. Any disagreement

[2] To obtain an idea of the impact of the sample composition on tagger performance, we evaluated the significance of the factors *text source* (*publisher*), *year of publication*, *topic* (Weiss 2005), and *country of publication* on the agreement between the three tools. Tukey HSD tests on the corresponding ANOVAs showed that there were significant correlations between all factors and the degree of agreement. However, taking into account the sample size, significant correlations were not surprising and the magnitude of the influences was rather small. We concluded that our virtual corpus is suitable for the comparison and that with respect to tagger performance, DeReKo is rather homogeneous (see Giesbrecht / Evert 2009 in contrast for web corpora).

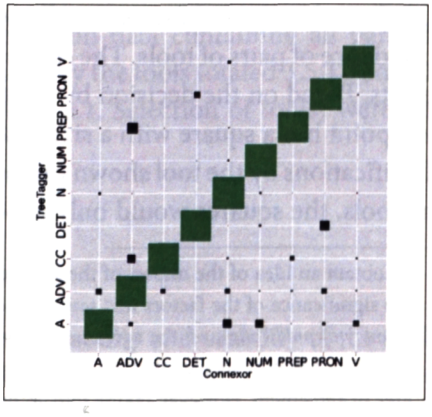
is depicted as a square outside the diagonals. For instance, as shown in Figure 1, tokens classified as adjectives by Xerox are often classified as nouns by Connexor.

B_9	Xerox	TreeTagger
A	ADJA, ADJA2, ADJA3, ADJD, ADJD2, ADJD3	ADJA, ADJD, VMPP
ADV	ADV, PTKANT, PTKCOM, PTKNEG, PTKSUP, WADV	ADV, PROAV, PTKNEG, PWAV
CC	COORD	KON
DET	ART	ART
N	NOUN	NE, NN
NUM	CARD, ORD	CARD
PREP	PREP, PTKINF	APPO, APPR, PTKZU
PRON	DEMDET, DEMINV, DEMPRO, INDDET, INDPRO, PERSPRO, POSDET, POSPRO, REFLPRO, RELPRO, REZPRO, WDET, WINV, WPRO	PDAT, PDS, PIAT, PIS, PPER, PPOSAT, PPOSS, PRELAT, PRELS, PRF, PWAT, PWS
V	VAFIN, VAINF, VINF, VMFIN, VVFIN, VVINF, VVIZU, VVPP	VAFIN, VAIMP, VAINF, VAPP, VMFIN, VMINF, VVFIN, VVIMP, VVINF, VVIZU, VVPP

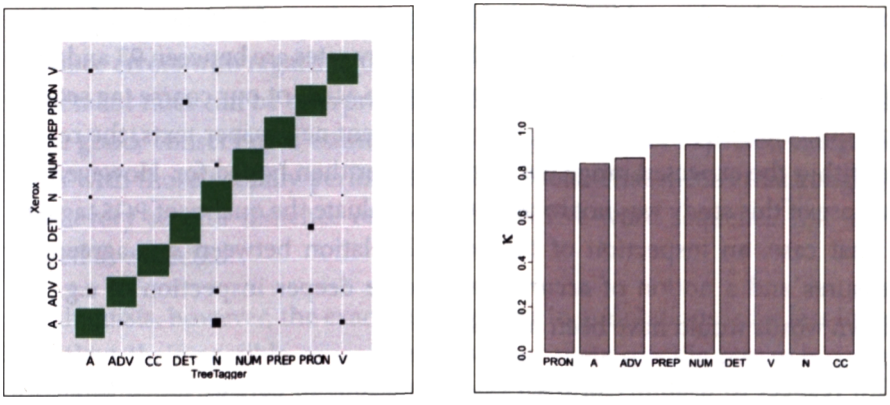
Table 2: Mapping to a coarse base tag set B_9



(a) Xerox – Connexor



(b) TreeTagger – Connexor



(c) Xerox – TreeTagger (d) by POS-tag

Figure 1: Tag correspondences and inter-tagger agreement

3.2 Assessment of reliability

Based on our coarse base tag sets, we also measured the overall agreement between the three tools with respect to single tokens. As shown in Table 3, the combination Xerox and TreeTagger had the highest percentage of agreement with 95.59% while TreeTagger and Connexor only agreed on 93.47% of the tokens and for the agreement of all three taggers the percentage drops to 91.57%. To account for by-chance matches and the rather small tag set, κ coefficients were calculated additionally, shown in the last column.

ts	tagger 1	tagger 2	tagger 3	%	κ
B_9	Xerox	TreeTagger		95.59	0.947
B_{26}	Xerox	TreeTagger		(94.40)	(0.935)
B_9	Xerox	Connexor		93.86	0.926
B_9	TreeTagger	Connexor		93.47	0.921
B_9	Xerox	TreeTagger	Connexor	91.57	0.931*
STTS	TreeTagger	Gold		93-98†	

Table 3: Inter-tagger agreement for single tokens (* Fleiss' κ, † reported accuracy)

There are only a few published reliability evaluations of POS-taggers for German and generally the reported overall accuracy rates are between 93 and 98 % (cf. Giesbrecht / Evert 2009), so that taking into account our coarse tag set and a corpus of mainly 'easy' but partially very recent newspaper texts, the results are within the expected range but slightly lower then hoped for. However, the purpose of this study was not to objectively evaluate the quality of POS taggers. In that case, an inspection of the general relation between our agreement measures and a notion of *accuracy* and some deeper inspection of e.g. unknown words would have been necessary.

With respect to usability of the POS annotations in corpus-based linguistic research, the proportions of full agreement on sentences, shown in Table 4, however, look somewhat alarming. Xerox and TreeTagger only agreed on every second sentence and all three taggers agreed on less than every third sentence.

ts	tagger combination	%
B_9	Xerox, TreeTagger	50.82
B_{26}	Xerox, TreeTagger	(42.16)
B_9	Xerox, Connexor	44.74
B_9	TreeTagger, Connexor	38.92
B_9	Xerox, TreeTagger, Connxeor	31.36
STTS	TreeTagger, Gold Standard	33.67-73.85*

Table 4: Inter-tagger agreement for whole sentences with mean sentence length = 15 tokens (*estimated based on reported accuracy)

4. Pitfalls

As already indicated in the previous section, a first possible pitfall for the linguistic exploitation of automatic POS annotations can arise from differences in part-of-speech taxonomies. A perfect mapping, at least to a fine-grained common tag set, was not straightforward and in addition to consulting the documentation required the comparison of samples. The linguist is confronted with a similar situation because he / she has to find out first how the categories used by the taggers relate to categories he / she has in mind.

The other, very obvious potential pitfall is of course that the tagging tools do produce errors with regard to their intended taxonomy. Whether the observed agreement rates can be interpreted as accuracy rates or not, they are probably quite good, conservative estimates for the agreement of expected categorizations with those actually performed by the tools. That means that roughly at least every second sentence or every fourth five-word sequence will not be tagged as expected.

For linguists, however, the exact rate of errors in annotations is far less important than their possible consequences within their specific research context. This will be the topic of the following three sections.

4.1 Error types

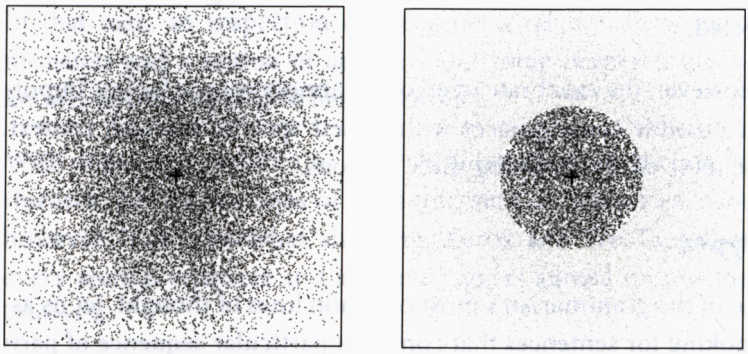
Mimicking one of the grammarian's most common lines of thought, let us assume we are looking for sentences that contain a particular sequence of parts of speech and the corresponding corpus query yielded 20 000 sentences as hits. If the accuracy of the result is roughly 75 % ($\approx 0.93^4$) and errors are distributed evenly, we are likely to have about 2 500 false hits among the 20 000: *false positives*. While this is bad news, the problem can be solved by sorting out the false hits manually. Much more problematic is that additionally we are likely to *miss* about 2 500 sentences we were actually looking for: *false negatives*. What if the unseen false negatives in fact contradict the findings based on the seen data?

As we have no access to such type II errors and consequently know nothing about them, there is a danger that without realizing we may end up not analyzing observations of language use but also the tagging tool, the theory and language model behind it, or possibly its imperfect implementation.

4.2 Error distribution

In this section, we present some considerations on the annotation error distribution from the point of view of linguistic research concerned with language variation in order to foresee and discuss possible dangers. Accordingly, our reasoning here is based on general qualitative assessments rather than hard quantitative evidence.

Let us assume we want to conduct research on language variation making use of a large collection of observations of language use events recorded in a corpus. Figure 2a gives an informal view of how the sample of language productions in our corpus is likely to be distributed around some *language core* (shown as black cross in Figures 2 to 7).

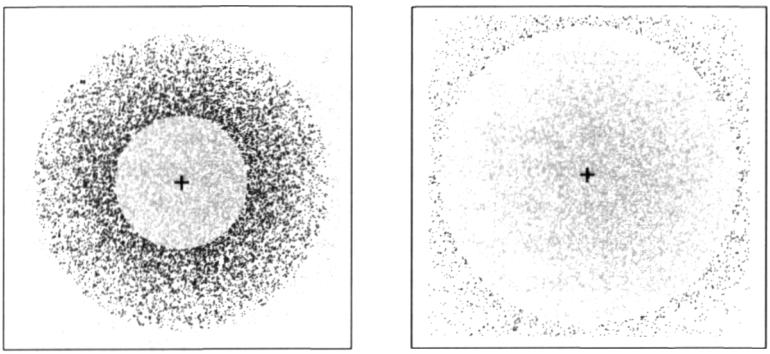


(a) Distribution of language phenomena around language core

(b) Focus on central phenomena

Figure 2: Schematic view of the distribution of language phenomena in a corpus and the focus of linguistic research

The further a point is from the centre of the plot in Figure 2a, the less familiar is the language phenomenon (LP) it stands for, e.g., it is the less frequent, the less standardized or conventionalized, the less uniformly distributed across areas, time, genres, topics, etc., the worse understood by general public, and so on. However, the central area need not represent the language core with respect to the language as a whole (which is the ambition of *representative general reference* corpora). Rather, it may refer to quite different realms of linguistic reality, it may be widely spread or tightly focused, e.g. on certain peripheral language phenomena (LPa), depending on specific sampling criteria used to build that corpus. Accordingly, we use the term *language core* to denote also the core LPa of any such – however skewed – sample of language productions throughout this paper. Since our assumed goal is to study language variation, we are likely to pay less attention to the central LPa covered by the corpus, cf. Figure 2b, and to concentrate on phenomena more distant from the corpus language core as shown in Figure 3a instead, possibly ignoring marginal phenomena highlighted in Figure 3b.



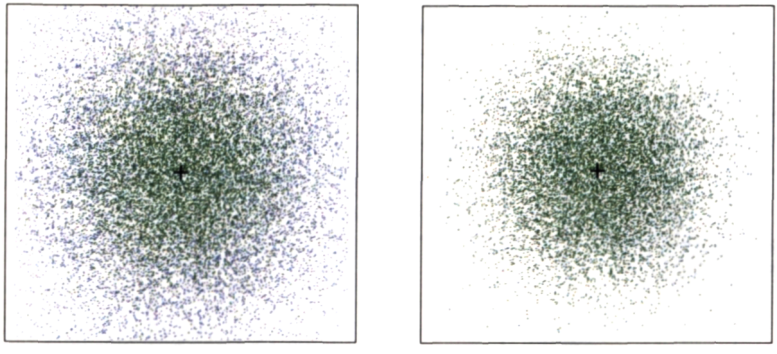
(a) Focus on language variation

(b) Focus on marginal phenomena

Figure 3: Focus of linguistic research

The intuitive notion of a language phenomenon used here is a very general one. It includes any observed language event which can sensibly be captured in terms of (any combination of) lexical, syntactic, semantic, stylistic, prosodic, phonetic, pragmatic, or other linguistic characteristics. Obviously, the dimensionality of the space spanned by these characteristics is considerably high. To clarify our line of thought, we use schematic two-dimensional plots in our examples nonetheless.

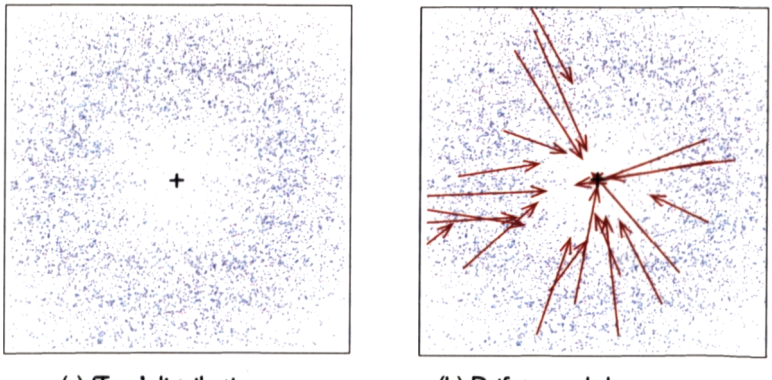
Let us also assume we have a piece of software that is capable of assigning such linguistic characteristics to observed language data, i.e., an annotation tool (tagger, parser, etc.), and we expect it to assist us in the process of exploring language variation. As is plainly evident, every annotation tool implements a notion of language core of its own, and, to keep our argumentation as simple as possible, we assume this language core is identical with that of our corpus. However, no annotation tool is perfect. While some LPa documented in our corpus are likely to lie within the linguistic scope of the annotation tool, others are not. It seems reasonable to expect that with increasing distance from the language core, the probability of a LP being within the scope of the annotation tool decreases. Figure 4a shows an assumed typical distribution of LPa within and outside the linguistic scope of an annotation tool.



(a) All phenomena (b) Within-the-scope phenomena

Figure 4: Distribution of language phenomena and the linguistic scope of the annotation tool (green: LP lies within the scope of the annotation tool; blue: LP lies outside the scope of the annotation tool)

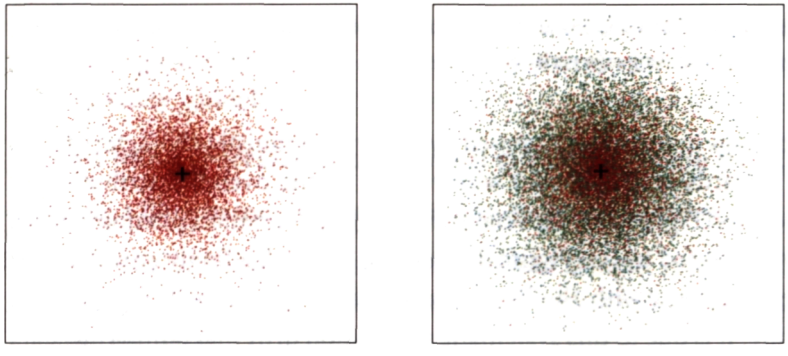
Let us now try to assess how accurate the annotation of these two groups of LPa is likely to be. Because the green dot phenomena (cf. Figure 4b) lie within the scope of the annotation tool, it is reasonable to assume that they are more likely to be annotated correctly than the blue dot phenomena shown in Figure 5a



(a) 'True' distribution (b) Drift towards language core

Figure 5: Language phenomena that lie outside the linguistic scope of the annotation tool and their distribution around language core

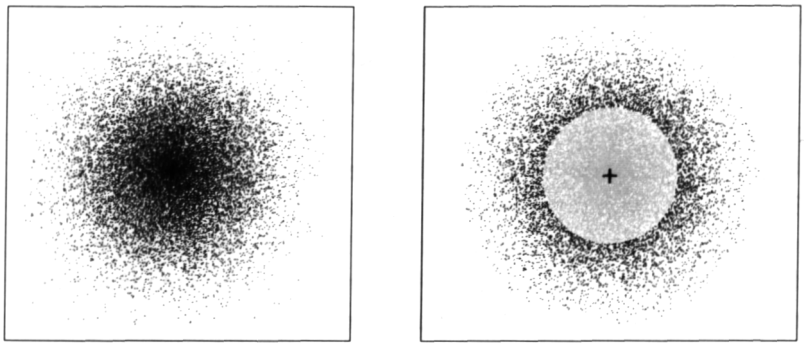
which lie outside the linguistic scope of the annotation tool. However, if the outside-the-scope LPa are more prone to annotation errors, the following question arises: can a sensible judgement be made about whether or not there is a systematic bias in the annotations actually attached to the outside-the-scope LPa by the tool as compared with their 'true' (or 'correct') linguistic characteristics? We suggest that random noise *plus a systematic drift* towards language core is introduced by the annotation tool into the data for the blue dot LPa, as indicated by red arrows in Figure 5b, provided the tool tends to assign some 'best guess' attribute to an LPa it is unaware of, which is what most annotation tools do.



(a) Annotations of outside-the-scope LPa (b) All annotations

Figure 6: Distribution of annotations around language core (green: the annotated LP lie within the linguistic scope of the annotation tool; red: the annotated LP lie outside the scope of the annotation tool)

Thus, the 'true' distribution of the LPa outside the scope of the annotation tool as shown in Figure 5a is mapped by the annotation tool to a biased distribution of – partially wrong – corpus annotation tags plotted as red dots in Figure 6a. In Figure 6b, this annotation tag distribution is plotted together with the distribution of the – predominantly correctly assigned – tags of the LPa shown previously in Figure 4b. Consequently, the resulting overall distribution (Figure 7b)



(a) Distribution of annotations (b) Focus on language variation

Figure 7: Distribution of annotations around language core and the focus of linguistic research concerned with language variation

of annotation tags in the space spanned by the 'true' linguistic characteristics is also biased towards language core with respect to the overall distribution of LPa in the corpus (Figure 2a). Especially the language variation phenomena we are actually interested in (cf. Figure 3a) tend to be misleadingly annotated as being more consistent with the linguistic scope of the annotation tool than they in fact are. It is due to this systematic annotation drift that our corpus data – if inspected indirectly, i.e., through the annotation layer – appear partially drained of linguistic variation (see Figure 7b).

4.3 Ways around

Unlike many technological (e.g. NLP or IR) projects, pure linguistic research is expected to be meticulously concerned with the theoretical status of its data and particularly of any annotation data included in language corpora. In linguistics, it is in general of crucial importance to construe annotation data as mere assessments (or 'opinions') of (human or automatic) annotators rather than as straightforward observations of language use. Consequently, these 'opinions' might sometimes turn out to be incomprehensible because of unclear or unfamiliar terminology (e.g., part-of-speech taxonomy), sometimes objectively wrong, and sometimes just grossly incompatible with one's own judgement.

A first remedy is to consider using more than a single source of such 'opinions', particularly when rare or non-standard linguistic phenomena are the object of research. However, this might still prove to be insufficient to avoid type II er-

rors. For instance, it is not unlikely that all tools are biased in the same or similar way, as sketched in the previous section. Thus, to increase the recall even more, it might be advisable to take into account not only those interpretations that were regarded as most plausible by the annotation tools but rather to consider all interpretations that were regarded as possible. While such an approach is supported by the current DeReKo annotation initiative, it is obvious that in extreme cases this might either still not be enough or that the resulting number of concurrent interpretations might render the use of the annotation data impractical.

There are no universal solutions, neither to the problem of uncertainty associated with relevant but unseen false negatives nor to the precision vs. recall trade-off. We suggest that a possible general strategy for using annotated corpora in linguistic research in a scientifically sound way is to adhere to the following 'safety' guidelines:

1) start with a general corpus query that aims at maximum recall;
2) apply a filter to remove the largest group of *false positives*;
3) cross-check manually (based on a random sample, if appropriate) if your filter has any undesired side-effects with respect to both *false negatives* and *false positives*;
4) adapt the filter and incorporate it into your corpus query;
5) repeat until *false positives* can be handled (e.g. filtered out) manually;
6) include the final query and a detailed error discussion in your publication.

Two of the authors have applied these guidelines in the course of two months' worth of research on German subject infinitival clauses with and without *zu*. A concise result of their endeavor has been published in Kubczak/Konopka (2008: 258f.).

5. Conclusions and Outlook

It is an inevitable fact that automatic linguistic annotations on a scale between *observation* and *interpretation* clearly have the status of interpretations if they depend on a norm, such as a POS taxonomy, which cannot be derived from the data. And in addition, annotations are typically erroneous with respect to such norms. Nevertheless, the use of, e.g., POS annotations can undoubtedly

make corpus query tasks easier. However, to achieve scientifically sound results based on automatic annotations, the adoption of a careful, potentially time-consuming strategy is indispensable. In general, such a strategy will be needed to avoid uncontrollable errors, as in search tasks most notably type II errors (false negatives). Such a strategy will typically involve an initial maximization of recall and many, eventually manual, iterations of filtering out false positives. In addition, to take account of error-prone interpretational status, the adoption of good practices from other sciences, as for example a strict experimental design and the discussion of possible sources of errors is particularly important.

While the necessity for such a careful modus operandi will never disappear, there is also ample room for improving the current annotation of DeReKo and its usability in linguistic research. Because the mission of the IDS is to conduct basic and applied linguistic research rather than to pursue language technology, we will not try to improve tagging tools ourselves, but, for instance, we consider providing a *maximum recall* and a *maximum precision* annotation layer to simplify the search and filter tasks respectively. Other techniques to enhance the benefits of annotations generated by third-party NLP tools in our research context might be to include auxiliary morphological analyzers to perform regular lexicon updates, or to apply the *ensemble of classifier* approach, where applicable.

References

Giesbrecht, Eugenie / Evert, Stefan (2009): Part-of-speech tagging – a solved task? An evaluation of POS taggers for the Web as corpus. In: Iñaki, Alegria / Leturia, Igor / Sharoff, Serge (eds.): Proceedings of the 5th web as corpus workshop (wac5). San Sebastian, Spain. Internet: http://purl.org/stefan.evert/PUB/GiesbrechtEvert2009_Tagging.pdf (last visited: 07 / 2010).

IDS = Institut für Deutsche Sprache (ed.) (1991-2009): COSMAS I / II Corpus Search, Management and Analysis System. Internet: http://www.ids-mannheim.de/cosmas2/ (last visited: 07 / 2010).

Koskenniemi, Kimmo / Haapalainen, Mariikka (1996): GERTWOL – Lingsoft Oy. In: Hausser, Roland (ed.): Linguistische Verifikation. Dokumentation zur Ersten Morpholympics 1994. (= Sprache und Information 34). Tübingen: Niemeyer, 121-140.

Kubczak, Jacqueline / Konopka, Marek (2008): Grammatical variation in near-standard German: A corpus-based project at the Institute for the German Language (IDS) in Mannheim. In: Štícha, František / Fried, Mirjam (eds.): Selected contributions from the conference Grammar and Corpora, Sept. 25-27, 2007, Liblice, Czech Republic. Prague: Academia, 251-260.

Kupietz, Marc / Keibel, Holger (2009): The Mannheim German Reference Corpus (DeReKo) as a basis for empirical linguistic research. In: Minegishi, Makoto / Kawaguchi, Yuji (eds.): Working papers in corpus-based linguistics and language education No. 3. Tokyo: Tokyo University of Foreign Studies (TUFS).

Schiller, Anne / Teufel, Simone / Stöckert, Christine / Thielen, Christine (1999): Guidelines für das Tagging deutscher Textcorpora mit STTS. Technical Report. Universität Stuttgart: Institut für maschinelle Sprachverarbeitung.

Schmid, Helmut (1994): Probabilistic part-of-speech tagging using decision trees. In: International conference on new methods in language processing. Manchester, UK, 44-49.

Tapanainen, Pasi / Järvinen, Timo (1997): A non-projective dependency parser. In: Proceedings of the 5th conference on applied natural language processing. Washington DC. Morristown, NJ: Association for Computational Linguistics, 64-71.

Weiss, Christian (2005): Die thematische Erschließung von Sprachkorpora. Technical Report. Mannheim: Institut für Deutsche Sprache. Internet: http://www.ids-mannheim.de/kl/projekte/methoden/te/te.pdf (last visited: 07 / 2010).